Radio-morphology: Parametric Shape-Based Features in Radiotherapy

by

Pranav Lakshminarayanan

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science and Engineering in Biomedical Engineering.

Baltimore, Maryland

December 2017

Abstract

In radiotherapy, it is necessary to characterize dose over the patient anatomy to target areas and organs at risk. Current tools provide methods to describe dose in terms of percentage of volume and magnitude of dose, but are limited by assumptions of anatomical homogeneity within a region of interest (ROI) and provide a non-spatially aware description of dose. The increased availability of medical image data in recent years has made possible the large-scale study of patient response to treatment. A practice termed *radio-morphology* is proposed as a method to apply anatomical knowledge to parametrically derive new shapes and substructures from a normalized set of anatomy, ensuring consistently identifiable features across a patient set. Radio-morphologic (RM) features are derived from a three-step procedure: anatomy normalization, shape transformation, and dose calculation. Predefined ROI's are mapped to a common anatomy, a series of geometric transformations are applied to create new structures, and dose is overlaid to the new images to extract dosimetric features; this feature computation pipeline characterizes patient treatment with greater anatomic specificity than current methods.

Examples of applications of this framework to derive structures include concentric shells based around expansions and contractions of the parotid glands, separation of the larynx into sections along the z-axis, and creating radial sectors to approximate neurovascular bundles surrounding the prostate. Compared to organ-level dose-volume histograms (DVHs), using derived RM structures permits a greater level of control over the shapes and anatomical regions that are studied and ensures that all new structures are consistently identified. Using machine learning methods, these derived dose features can be coupled with patient data and clinical assessments to uncover dose dependencies of inter- and intra-organ regions. Radio-

morphology is a valuable data mining tool that approaches radiotherapy data in a new way, improving the study of radiotherapy to potentially improve prognostic and predictive accuracy.

Readers:

Dr. Todd R. McNutt

Dr. Russell H. Taylor

Dr. Jeffrey H. Siewerdsen

Acknowledgments

I would like to thank my advisors, Todd McNutt and Russell Taylor, for their guidance on this project, and Jeffrey Siewerdsen for his review and feedback on this thesis.

I would also like to thank Scott Robertson, Wei Jiang, Sierra Cheng, Xuan Hui, Peijin Han, Ilya Shpitser, Sauleh Siddiqui, Harry Quon, Michael Bowers, Joseph Moore, and Alex Mathews for all their help, and the Radiation Oncology Institute for funding this research.

Finally, I would like to thank my parents, for all these years of love and support.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Thesis statement

Parametric shape-based models of dose in radiotherapy provide valuable insight into the quality of treatment plans, better understanding about the radio-sensitivity of human anatomy, and can help tailor treatments to prevent harmful post-treatment side-effects.

1.2 Big data and radiotherapy

"Big data" has become increasingly prevalent in healthcare; the defining characteristics of big data (volume, variety, velocity, and veracity) allow for new knowledge that is only made possible with information at this scale [1]. Big clinical data refers to the quantifiable values that describe a patient's history, their body, their treatment, and their clinical outcomes. More specifically within the scope of radiotherapy, this data includes *medical images* of patient anatomy, *dose maps* that detail how radiation is to be delivered to a patient, and *quantitative clinical assessments* that describe a patient's condition due to the treatment [2].

Clinical data also poses its own challenges, such as a variety of data types, missing data over time, error in measurement, and lack of resolution based on the assessment tools; however, these limitations can often be overcome using a larger volume of data. Thus, to make use of the large scale of data, it is necessary to have tools and methods that are robust enough to efficiently and accurately extract meaningful information. Applying machine learning methods to big clinical data opens the door to information that was previously unavailable, such as the effect of radiation therapy on human physiology. Ultimately, the role of big data in radiotherapy is to provide a means to gain new knowledge into clinical practices and develop models to improve future treatments.

1.3 Current practices

Before presenting a new method, it is necessary to understand the types of data available and the limitations of current practices. The major treatment-related data types that are considered in this thesis are *medical images* and *dose maps*. Medical images provide a representation of the patient anatomy, making it possible to identify regions of interest (ROIs), such as tumors and organs at risk (OARs). Computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and single photon emission computed tomography (SPECT) are common imaging modalities in radiotherapy, each providing their own benefits, such as visibility of different types of tissue and visualization of physiologic processes [3]. Dose maps are three dimensional images that detail the planned delivery of dose to a patient. By overlaying dose grids onto a patient's images, it is possible to study the distribution of dose to specific parts of the patient's anatomy.

Dose volume histograms (DVHs) are the current standard for characterizing radiation dose across a patient's anatomy. While this method provides an intuitive and easily understandable method to describe the dose in a region, DVHs pose two major limitations: loss of spatial information, and containment to pre-defined structures in pre-treatment imaging.

The DVH decreases the dimensionality of a three-dimensional dose distribution to a one-dimensional histogram describing what percentage of a structure receives a given amount of radiation (example shown in Figure 1). These plots are based on an assumption of functional homogeneity, that is to say, all parts of an organ exhibit the same radio-sensitivity and contribute equally to function. While DVHs are easily interpretable and provide an accurate summary of the dose delivered to a structure, this calculation does not include any spatial information or a way to encode different radio-sensitivity or functional importance [4].

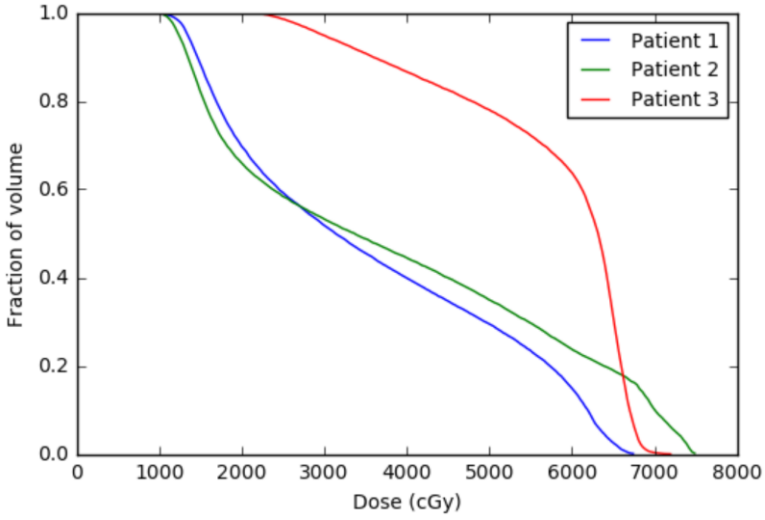
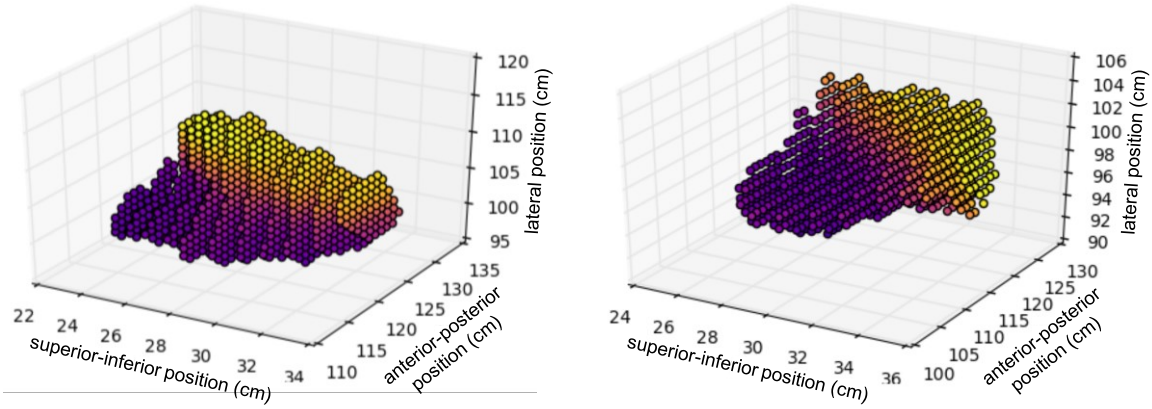


Figure 1: Sample DVHs of three parotid glands



(a)
Dose map for Patient 1

(b)
Dose map for Patient 2

Figure 2: Dose maps of right parotid structures

Figure 1 shows an example of DVH curves describing the dose to the right parotid gland in three patients. While patients 1 and 2 received similar amounts of dose compared to the patient 3, there is no information to what parts of the structure received the radiation, meaning that the underlying dose maps could look vastly different between the three patients.

The dose maps in Figure 2 show the spatial distribution of dose on the ROIs for patients 1 and 2. In this representation, it is possible to see the clear differences in the spatial dose distribution that were not made evident in the DVH curves.

Another limiting characteristic of the use of DVHs in dose distribution analysis is the scale of structures available from pre-treatment imaging. The contours that define the regions are biased towards structures that are visible within the pre-treatment imaging modality and are important to treatment planning. For example, in the head and neck (HN) region, this could include the left and right parotid glands, but would not distinguish intra-organ areas, such as regions close to the salivary ducts. Similarly, the prostate and rectum are commonly contoured, but not the neurovascular bundles surrounding the prostate. This lack of anatomical specificity thereby reduces the information that can be extracted from a dose distribution, and thus, limits the utility of organ level DVHs.

1.4 Goals and research objectives

Having described the limitations of solely using organ level DVHs as an instrument to describe dose distributions, several goals can be identified for the solution put forth. The primary goal is to be able to *generate spatially representative dose characteristics* for machine learning applications to discern relationships between dose to an organ and function or change in patient outcomes. Thus, it must be possible to *develop methods to produce parametric models* of the anatomy. Once the parameters used to derive the dose models have been defined, it is then possible to tune the dose features that are generated within machine learning applications to maximize predictive power in order to discover the functional relationships within anatomy and better understand the response to radiation.

1.5 The concept of radio-morphologic dose features

The concept of *radio-morphologic (RM) features* is proposed as a method to parametrically describe the spatial dose distribution within anatomic structures. These features can range from voxel-based analysis, where dose to individual elements (voxels) on an image are sampled [4], to shape based features, which are structures derived from an ROI using geometric transformations. The goal of this method is to overcome the limitations of current techniques by providing methods that create anatomically and spatially aware representations of radiation to a patient.

By creating a standard pipeline for feature generation, it is possible to derive consistently identifiable characteristics of an image. By composing geometric transformations to images on a standard reference frame, new structures can be extracted. Then, radiation dose can be mapped onto the new structures to extract new dose characteristics. The following sections describe the RM feature generation pipeline, a specific implementation, and applications of this framework for anatomy approximation and prediction of post-treatment, radiation induced toxicities.

2 Methods: The Feature Generation Pipeline

The RM feature generation pipeline consists of three steps: *anatomy normalization*, *image transformation*, and *dose feature extraction*. Anatomy normalization is the process by which anatomical structures are deformed to a standard coordinate frame, allowing points on an ROI to be compared across a patient population. Then, geometric transformations are applied to manipulate a structure and derive new shapes. Finally, radiation dose is mapped onto the new shapes, and dose features are extracted. Figure 3 shows a general representation of the feature generation pipeline, broken down into the three major steps.

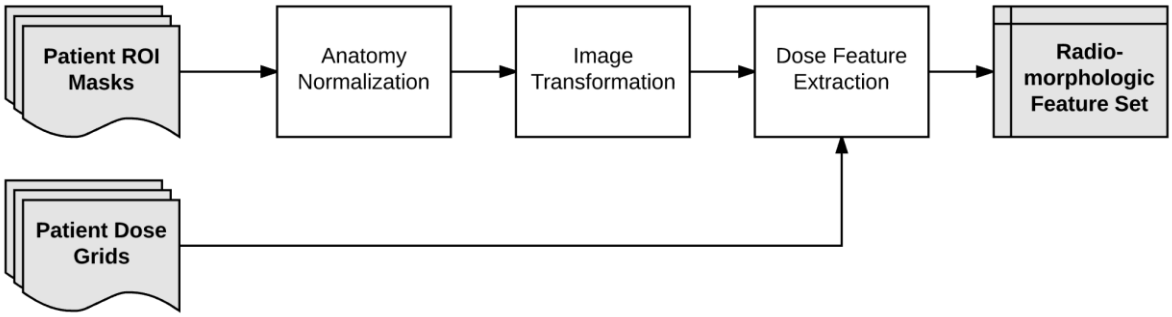


Figure 3: General feature generation pipeline

2.1 Anatomy Normalization

2.1.1 The need for anatomy normalization

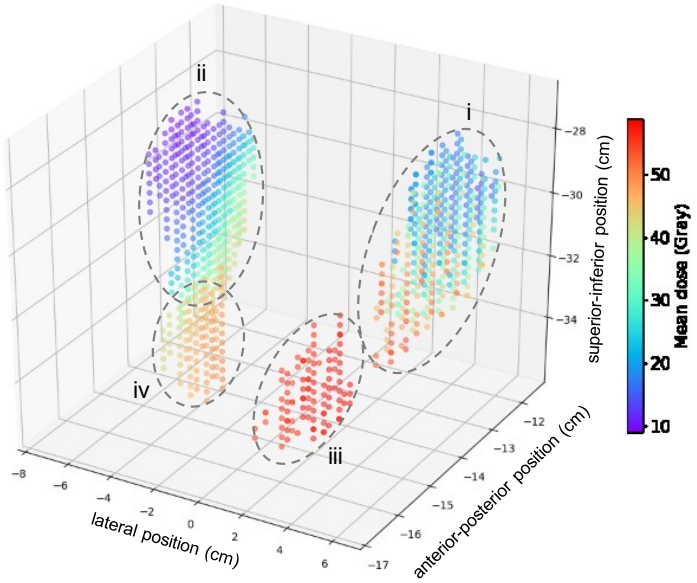
Anatomical structures in medical images are inherently different from patient to patient, making it impossible to consistently characterize them across a patient population; thus, it is necessary to find a way to manipulate all structures on the same geometric reference frame. To do so, the first step of the pipeline is *anatomy normalization*. Using a reference image, or

anatomical atlas, each patient's anatomy is deformably registered onto the reference image. Any suitable deformable registration algorithm can be selected for this step, based on the type of the data available. In addition to the variety of imaging modalities (CT, MRI, PET, SPECT), anatomical contours can be extracted to produce binary masks of ROIs; contours and ROI masks are further examined in following sections as the main representation of patient anatomy. To extract information across these varying image types, deformable registration algorithms can be applied to normalize the anatomy in an image [3]. While the application of deformable registration to image sets includes some error based on the specific algorithm used and the level of variation across the set of patients, the result of this normalization step is a set of images where patient anatomy is better aligned across the population, allowing for more consistent feature extraction.

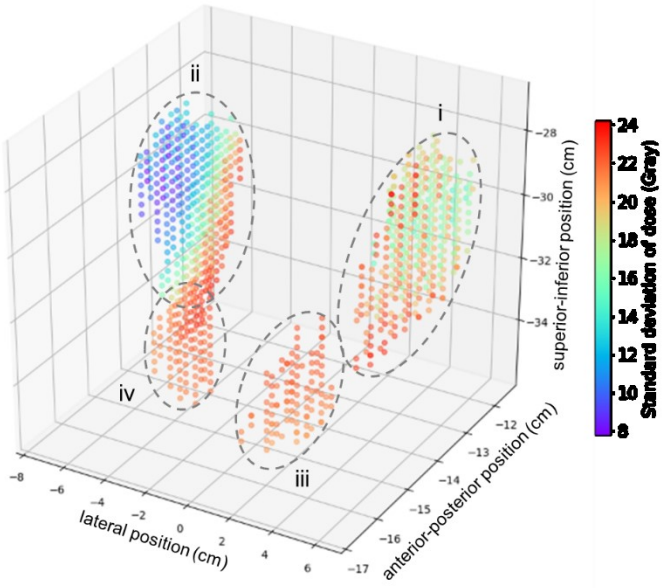
Beyond inter-patient anatomy normalization, it is also possible to normalize structures by laterality. Rather than considering regions as left- and right-lateral, it is possible to label anatomy as ipsi- and contra-lateral to the radiation target. With parallel anatomy, treatment dose varies not by left or right location, but rather by location with respect to the treatment target. Pairs of symmetric ROIs, such as the parotid and submandibular glands, can be mapped to their corresponding sides by mirroring one across the medial axis and registering the pair of ROIs to each other. In doing so, any derived shapes or features are defined by laterality to the target area. Normalizing laterality with relation to the target area is essential to being able to discern the relationships between dose delivery and radiation induced outcomes.

2.1.2 Application: Aggregate dose statistics

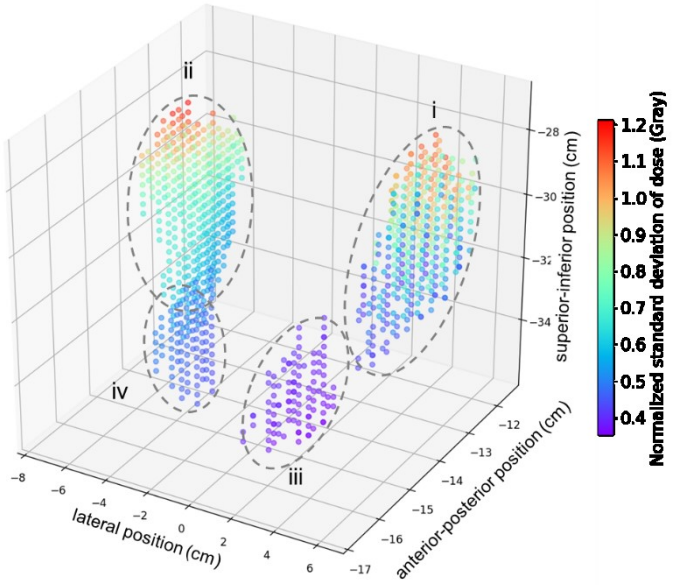
Anatomy normalization allows for insight into treatment patterns across a patient population. By registering a set of dose images to a common reference frame, it is possible to accumulate dose statistics and better understand trends in treatment planning. Figure 4 shows dose statistics generated from a cohort of 427 HN cancer patients. The parotid and submandibular glands were normalized using coherent point drift deformable registration (further explained in Section 3.3) [5]. Laterality regularization was also performed; in each visualization in Figure 4, the contralateral glands appear on the left and the ipsilateral glands appear on the right. Figure 4(a) shows the average dose distribution across the patient set, demonstrating that the ipsilateral submandibular receives the highest dose, and the superior region of the contralateral parotid receive the lowest dose. Figure 4(b) shows a similar pattern for the standard deviation of dose: the dose to the contralateral submandibular and the medial face of the contralateral parotid have a high variance across the patient set. Figure 4(c) computes the coefficient of variation, which is the standard deviation from (b) normalized by the mean dose distribution from (a). The aggregate dose statistics presented in Figure 4 are revisited in Section 4.2.1 alongside voxel based analysis for clinical outcome prediction. Thus, the anatomy normalization step adds consistency to the generation of dosimetric features across a patient population.



(a) Average dose distribution



(b) Standard deviation of dose



(c) Coefficient of variation. Computed as (b)/(a)

Figure 4: Dose statistics from CPD normalized parotid and submandibular glands over a set of 427 patients

In each figure, four structures are identified:

i. Ipsilateral parotid gland
ii. Contralateral parotid gland
iii. Ipsilateral submandibular gland
iv. Contralateral submandibular gland

2.2 Shape Transformations

2.2.1 Characterizing Shape Transformations

Shape transformations are the foundation of radio-morphology, whereby images and structures are morphed and manipulated to generate new structures. These derived structures are motivated by knowledge of the anatomy to transform existing, pre-defined structures into new, anatomically significant structures. The two basic classes of transformations in this step are *scaling* and *partitioning*.

Scaling transformations consist of expansions or contractions of a structure in any direction. Expansions are commonly used to define an ROI beyond just the surface of a single structure, allowing it to encapsulate critical regions that may be adjacent to the original ROI or allow a margin for error that could occur during treatment setup. By contrast, contractions can be used to scale structures down and ignore regions near the surface. *Partitioning transformations* use axes and planes to carve up one structure into multiple substructures. By manipulating the coordinate frame on which the points lie (Cartesian, cylindrical, etc.), an ideal partitioning method can be selected for the anatomy.

To produce the correct set of derived shapes, a set of transformations can be composed and applied in series. While a single class of transformations can only produce a small set of shapes, a composition of transformations can result in more complex and specific derived structures. Figure 10 in Section 4.1 shows an example where several transformations were composed to create a region that approximates a region of neuro-vasculature surrounding the prostate. The following sections further describe examples of image transformations with visualizations of their applications to certain anatomy.

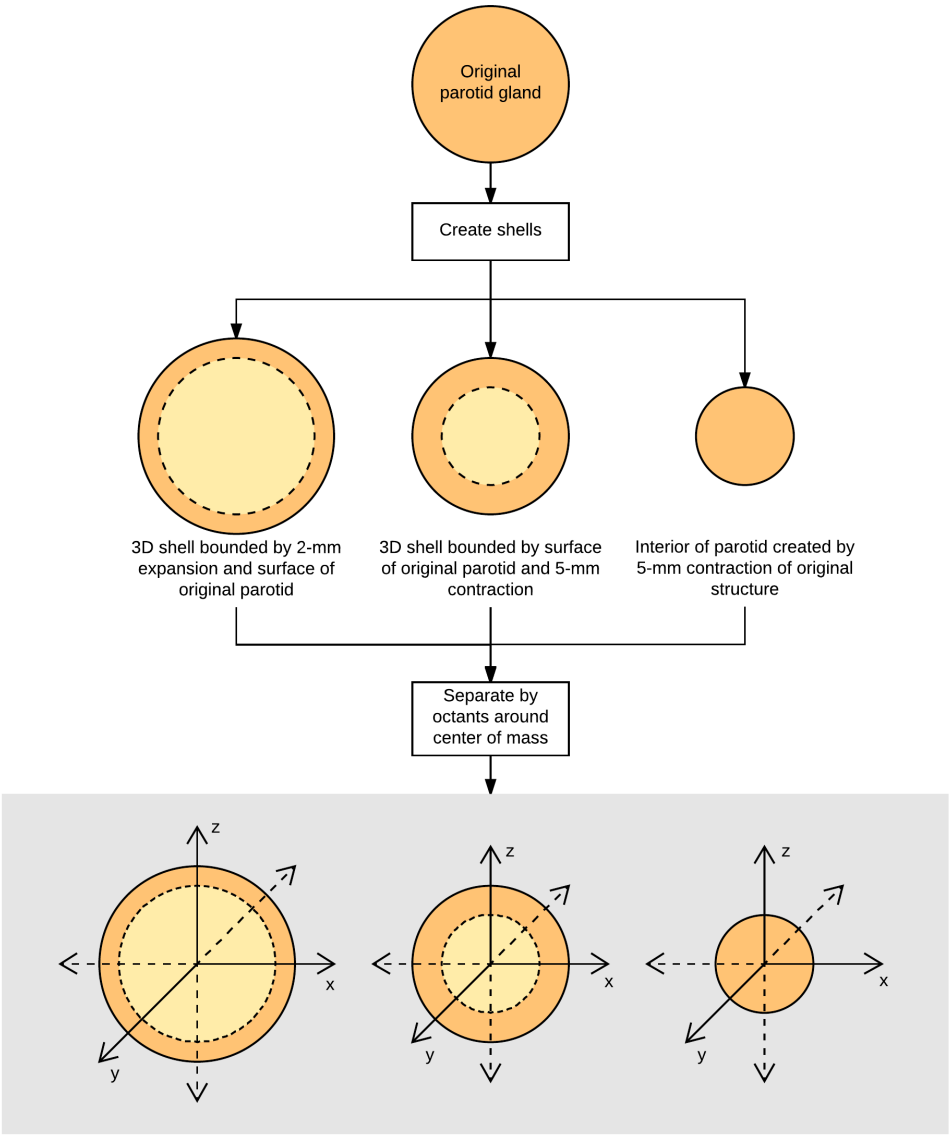
All transformations are mathematically defined with a set of parameters, for both consistent, repeatable definitions and efficient modification with parameter tuning. The latter is vital for feature exploration where the derived shapes can be iteratively modified; for machine learning applications using dose features of structures that are produced via shape transformations, this tunability allows optimal features to be found. In addition, the parameterization of human anatomy adds a flexibility to the data that is not afforded only using predefined ROIs. While it would be ideal that anatomy is finely contoured, the shapes derived from image transformations can be used as a surrogate to approximate the anatomy that is not visible or known.

2.2.2 Shells, Slices, and Octants

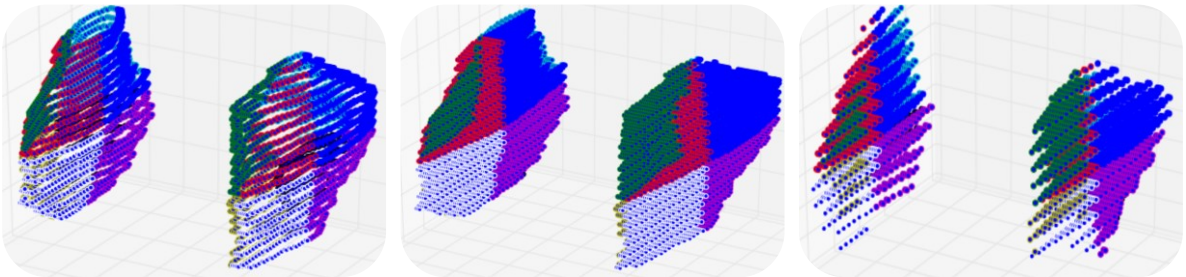
Shells are shapes that result from a composition of expansions and contractions. A set of expansion and contraction parameters can be used to create new shapes. Then, applying XOR Boolean logic to exclude structures with a smaller radius produces shells. Shells are applicable shapes for anatomy with radial properties or in cases when trying to capture extra-organ regions.

Slices and *octants* are simple partitioning transformations in Cartesian space. Slices can be defined by an axis to separate a structure into sections along that axis. Slices are well suited for axial structures, such as the esophagus or larynx, where dose could vary greatly along their axis. Octants perform a similar partition, but rather than being defined by one axis, are produced by separating along the x, y, and z axes centered at a given point, most often the center of mass of a structure. This method of partitioning produces a systematic spatial representation of a structure, making it possible to observe the anterior/posterior, superior/inferior, and medial/lateral regions of an organ.

Figure 5 shows (a) a composition of shells and octants and (b-d) the application of this set of transformations to the parotid glands to derive many substructures. First, shells were generated with bounds at a 2-mm expansion, the original surface of the glands, and at a 5-mm contraction. This transformation results in 3 new structures which are based on scaling transformations. Then, octants were generated by centering a Cartesian grid at the center of mass of each shell. The planes defined by the axes separate each shell into 8 substructures. This composition of transformations demonstrates how a single structure can parametrically be deconstructed into many substructures, in this case 24 new shapes per parotid, and capture information about all regions of the structure, such as the core, the surface, and the extra-organ space. The visualizations in Figure 5(b-d) show the color-coded octants for each shell.



(a) Cartoon representation of shells and octants transformation



(b) Shell bounded by a 2-mm expansion

(c) Shell bounded by surface of parotid glands

(d) Shell bounded by a 5-mm contraction

Figure 5: Parotid glands transformed by a composition of shells and octants

2.2.3 Sectors

Sectors are a partitioning transformation that uses a change of coordinate system to produce a new variety of shapes. By converting a Cartesian image to polar coordinates, radial sectors can be computed that are bounded by angles around a central point. This method of partitioning is more flexible than octants, making it possible to define regions with greater specificity than Cartesian octants. Applications of sectors are shown in section 4.1, where this partitioning transformation is used to approximate regions of neuro-vasculature surrounding the prostate.

2.3 Dose Feature Extraction

2.3.1 Characterizing dose to derived regions

Once new structures are computed, the final step is to extract dosimetric features that characterize the radiation dose across each substructure. When mapping dose grids onto the set of masks, it is important to account for differences in voxel size between the shapes and the dose grid with tri-linear interpolation. The output of the dose alignment step is a *dose map*, shown in Figure 2, at the resolution of the anatomical masks, where each voxel holds a dose value. With these dose maps computed, it is now possible to extract dosimetric features to each structure at a finer, more physiologically relevant resolution than whole organ DVHs.

2.3.2 Dosimetric features and dose-volume histograms

The purpose of the dose feature extraction step is to compute a set of values that describes the dose to the regions derived by the shape transformations. This step is an exercise in dimensionality reduction, where a three-dimensional representation of an anatomical structure is projected onto fewer dimensions (in the case of a DVH) or sampled at a coarser resolution to reduce the number of points characterizing the dose for voxel-based analysis [4].

While dose-volume histograms were previously criticized for their inability to encode spatial information, the DVH is reintroduced here as a viable tool for dose feature extraction. Following the normalization and transformation steps, the resulting structures are anatomically motivated and represent the patient anatomy at a finer resolution than organ level contours. At this level, computing a DVH for each derived structure balances the loss of spatial information from a DVH with the idea that each structure itself carries enough spatial importance on its own. That is to say, since the derived shapes encode a more specific spatial positioning and physiological meaning within an organ, a DVH calculation is a viable method to extract dosimetric features.

For an even finer resolution without the loss of spatial information, voxel level dose sampling can be used to reduce the number of dose features in a structure. This method changes the resolution of the dose mask by sampling points across the structure, decreasing the number of points overall, but without losing any dimensionality to maintain a spatial representation of dose to the derived structures. The normalization step ensures that the anatomical significance of a sampled voxel is consistent across a patient population, meaning that voxel sampling is an applicable method to characterize the dose distribution [4].

Figure 6 shows a complete description of the processing workflow shown previously. Each phase of the pipeline (anatomy normalization, image transformation, and dose feature extraction) has been further expanded to show the inputs, methods, and outputs. Figure 7 provides pseudocode for the RM pipeline, with generalized function descriptions for each step. While this section provided a conceptual definition of RM features, Section 3 describes a programming implementation to make use of clinical data and Section 4 presents various applications of this methodology.

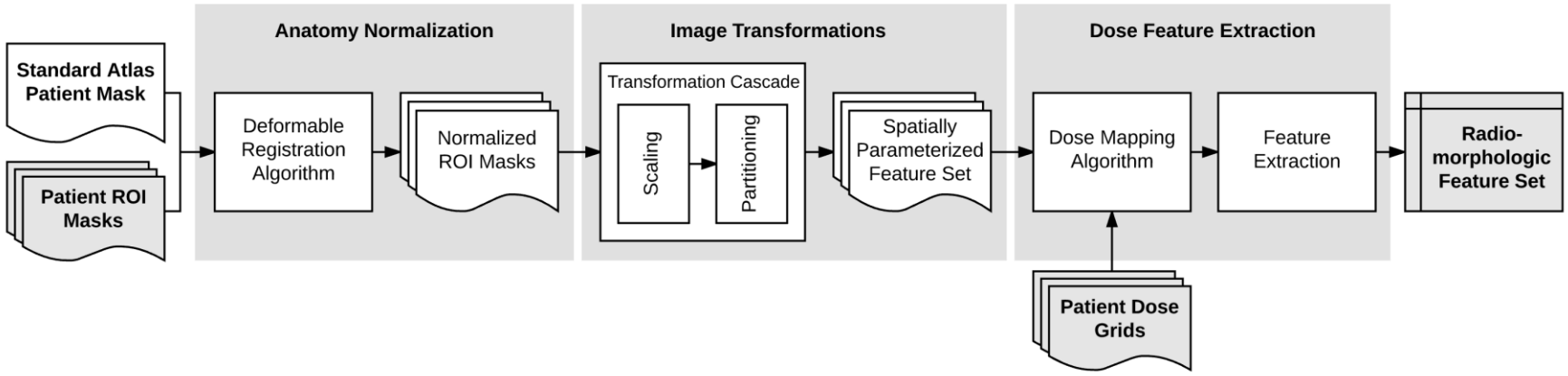


Figure 6: Feature generation pipeline

Anatomy normalization: Register patient masks to an anatomical atlas

```
function register(patient_img, atlas):
    '''
    This method can contain any registration algorithm
        that suits the data types of the patient mask and atlas
    Arguments:
        patient_img:  patient ROI image
        atlas:        atlas ROI image
    Returns:
        norm_img:  patient ROI mask in atlas coordinate frame
        V:         transformation parameters for registration
    '''
    norm_img, V = registration_algorithm(patient_img, atlas)
    return norm_img, V
```

Extract dosimetric features

```
function calc_dose_features(dose_map):
    '''
    A variety of dosimetric features can be extracted from a dose map
    For example, DVH values, minimum, maximum, and mean dose
    Arguments:
        patient_img:  patient ROI image
        atlas:        atlas ROI image
    Returns:
        dose_feats:   dictionary of dosimetric features
    '''
    dose_feats = {
        'min':  min(dose_map.data),
        'max':  max(dose_map.data),
        'mean': mean(dose_map.data),
        'dvh':  compute_dvh(dose_map.data)
    }
    return dose_feats
```

Figure 7: Pseudocode for RM pipeline

Create RM Features

```
import transform

# Create the sequence of transformations
assign transformation_sequence

given patient_img
given dose
given atlas

# Normalize anatomy
norm_img, V = register(patient_img, atlas)

# Perform the series of shape transformations
tf_imgs = mask
for tf in transformation_sequence:
    tf_imgs = apply(tf_imgs, tf)

# Map dose to shapes
#   For this step, masks and dose must be in the same coordinate system
for tf_img in tf_imgs:
    # Apply the inverse transformation
    #     to return the derived shape to the original coordinate system
    tf_roi = apply(tf_img, inverse(V))

    # map_dose_to_image is a function that varies with the data format
    dose_map = map_dose_to_image(tf_roi, dose)
```

Figure 7 (cont.): Pseudocode for RM pipeline

3 Implementation

3.1 The Oncospace database

The Oncospace database at The Johns Hopkins Hospital, Department of Radiation Oncology in Baltimore, MD, is a Microsoft SQLServer database that stores clinical data about radiotherapy procedures at this institution. The database was designed to store a variety of data, including patient histories and demographics, binary masks of ROIs, dose grids, clinical assessments of toxicities and quality of life (QoL), and quantitative information about diagnosis and disease progression [6]. By gathering data from a variety of sources, such as the treatment planning system and from the oncology information system (OIS) in the clinical workflow, it is possible to accumulate a breadth of data for many patients and store it in a structured database for consistent access. The Oncospace database served as the data source for the development of the radio-morphology framework. The specific data types drove the design choices made when developing a codebase to produce RM features.

Figure 8 presents a simplified schema of the Oncospace database, showing tables that are the most relevant to the RM pipeline (Not shown are tables for patient history, protected health information (PHI), clinical events, and pathology data) [7]. The *PatientRepresentations* table contains an ID for a specific patient and the geometric parameters defining the patient's CT images, such as the physical coordinates of the start of the image grid, size of the image in x, y, and z dimensions, and size of the voxels in physical coordinates. For each patient representation, the *RegionsOfInterest* table stores a set of binary masks of the anatomy, contoured from the patient's CT image. Storing the anatomical masks as run-length encoded binary images allows for small data size, fast queries, and preserves the shape and structure of the ROIs. To examine dosimetric features, the *RadiotherapySessions* table stores dose grids

that represent the distribution of radiation to be delivered during treatment. Note that the dose grids have their own coordinate system, meaning that when computing dose maps, it is necessary to convert both the ROI masks and dose grids to physical coordinates and interpolate dose values to the resolution of the ROI.

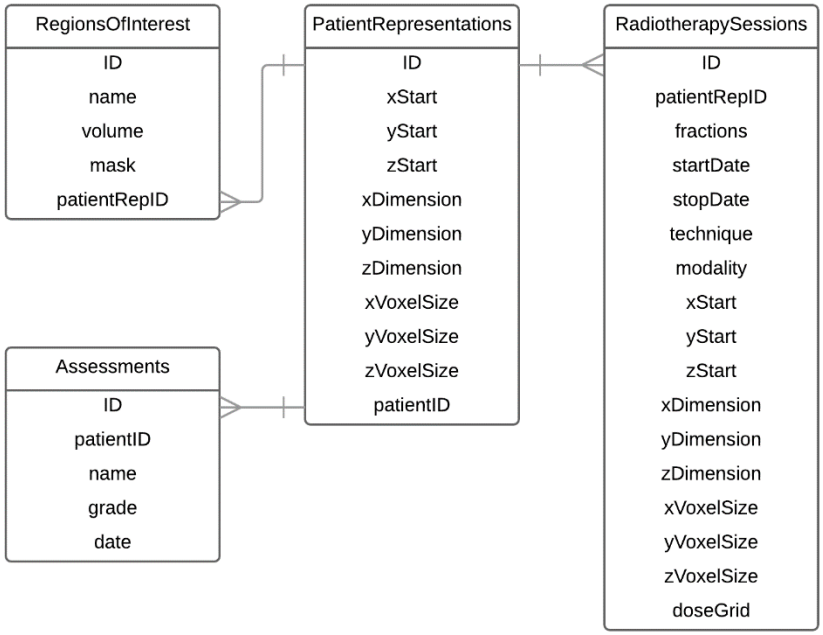


Figure 8: Simplified schema of Oncospace database for the RM pipeline

3.2 `oncotools`: A Radio-morphology library

A Python library, named `oncotools`, was developed to interface with the Oncospace database and construct the pipeline for RM feature generation. The library was written in Python 2.7 using the *SciPy* stack of libraries and tools, including `numpy` for data storage and matrix operations, `matplotlib` for visualization, `pandas` for data interaction, and `IPython` in Jupyter Notebooks [8] [9] [10]. The following sections describe various modules within the developed codebase.

3.2.1 `OncospaceConnect`: database interaction

The `OncospaceConnect` module contains classes that connect to the Oncospace database and carry out SQL commands from within Python; the `Database` class performs these function by wrapping two `pyodbc` functions, `connect()` and `execute()` [11]. Upon initialization, the `Database` class validates the connection details with the specified data source. A collection of common queries is built into the `Database` class, but any custom query can be passed in to the `Database.execute(query)` function as a string. The results of the SQL query are returned as an instance of the `Results` class, which is a structured representation of the query results as a two-dimensional array, along with metadata regarding the column names and number of rows in the output. Figure 9 shows a sample code snippet to connect to the database and perform a query.

Using the `OncospaceConnect` module

```
from oncotools.OncospaceConnect import *

db = Database('{SQL Server}', 'host-url', 'database-name', 'username', 'password')

query = '''
    SELECT ID, patientID
    FROM PatientRepresentations
'''
results = db.execute(query)
```

Figure 9: Sample code to connect to the Oncospace database and perform a query

3.2.2 The `Feature` base class

The `Feature` base class is an abstraction of the concept of an RM feature. Every `Feature` class is based on an ROI mask and a dose grid. The `process()` method carries out a predefined set of geometric transformations, creates a dose map, and extracts dosimetric

values. The purpose of abstracting the `Feature` class is to create a simple basis for any RM feature calculation; this design adds flexibility by making it possible to create many different types of features using different sets of transformations, but also enforces a consistent class structure. Using the functions in the `transform` module to perform shape transformations, new subclasses can be created that extract different RM features from an ROI mask. Appendix A contains excerpts from the Python codebase for the `Feature` base class and one implementation, the `VolumetricFeature`, which utilizes the shell transformation, described in Section 2.2.2.

3.3 Coherent point drift for anatomy normalization

Coherent point drift (CPD) is presented as one deformable registration algorithm that can be used for the anatomy normalization step with the types of data from the Oncospace database. CPD is a probabilistic method that is used to register sets of points (such as ROI masks) rather than gray-level images (CT, MRI, etc.). The registration algorithm is a maximum likelihood estimation that uses motion coherence as a constraint for the velocity field, ensuring that the structure of an ROI, or the relationship between points in an image, is maintained [5]. This algorithm is suited for structures represented as three-dimensional point clouds and simultaneously solves for the transformation and the point correspondence, meaning that this procedure can operate on segmented images without the need for landmarks, unlike many other deformable registration algorithms.

The Oncospace database stores three-dimensional binary masks of each contoured ROI, rather than gray level images. This design choice allows for fast query and contains sufficient information regarding the shape and structure of the anatomy. These masks were converted to Cartesian point clouds and then used as input to the CPD method to map all patient

structures to a common geometric reference frame. CPD excels in situations with unequal or missing data; in this scenario without image landmarks and a naturally varying number of points between patients based on size of the anatomy, CPD is thus an effective method to be used for the anatomy normalization step [5].

CPD can also be applied for laterality normalization. After mirroring the anatomy from one side across the medial axis, CPD creates correspondences between points from the pair of structures. The resulting normalized structures permit feature extraction that is fully defined by the anatomy and shape of each organ.

4 Applications

Having explained the RM feature generation pipeline, it is now possible to understand the applications of this procedure in a clinical research setting. Two applications are presented that highlight different steps in the pipeline and demonstrate the value of radio-morphology in data science for radiotherapy.

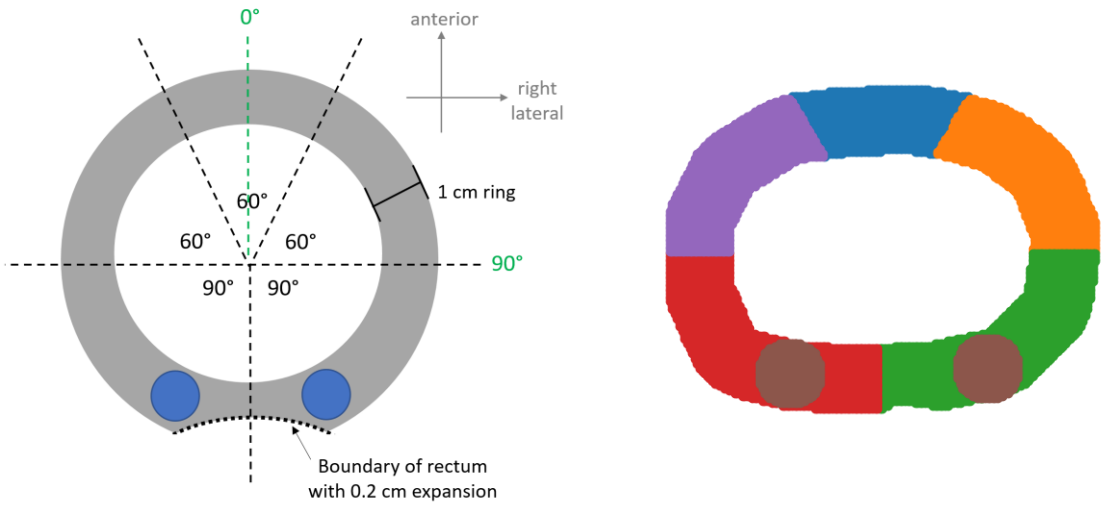
4.1 Approximation of non-contoured regions

Current dosimetric practices are often restricted by the use of predefined contours. The availability of contours of a structure is based on factors such as visibility on the imaging modality, value to design of a treatment plan, and physician effort to contour that structure. In any case, there will be certain anatomy that goes un-contoured, and recovering the desired areas can be time consuming and impractical at a large scale. However, the RM pipeline provides a method to approximate non-contoured regions for a patient population using anatomy normalization and shape transformations. By registering all anatomy to a common geometric reference frame and applying a set of transformations, new shapes can be created that represent previously unidentifiable regions. One example of this application is the approximation of the neurovascular bundles (NVBs) around the prostate.

The neuroanatomy surrounding the prostate is linked to continence and potency, and as a result, should be spared in oncologic treatments. Common nerve sparing techniques assume the location of the NVBs to be symmetrically located posterolaterally, but studies have also identified that minor bundles are spread more widely along the anterior face of the prostate [12]. This variability and uncertainty in the location of the NVBs makes them difficult to contour, but more suitable to approximate an area where the major NVBs and other minor neuro-vasculature may be found; in predictive studies to understand radiation induced changes

in sexual function, dose to all neuro-vasculature surrounding the prostate can be considered by producing these approximate regions.

A study was performed to create a spatially parameterized feature set that approximates a region containing the major NVBs and minor neuro-vasculature around the prostate. A training set of 115 manually contoured NVBs were used to validate the transformation parameters. In addition to encapsulating the major bundles that are located posterolaterally, other regions were defined around the surface of the prostate to approximate the regions where the minor bundles would be found. The spatial regions, shown in Figure 10, were defined as a linear combination of given contours, followed by a sector partitioning. First, the prostate was expanded by 1 cm in the x-y plane. Subtracting the original prostate area leaves a 1 cm thick shell around the surface of the prostate. Then, to prevent overlap of the new shell into the rectum, an area defined by a 0.2 cm expansion of the rectum was subtracted. Finally, with the remaining shell, sectors were calculated about the superior-inferior axis bounded by angles of 30°, 90°, 180°, 270°, and 330°. The resulting regions are shown colored in Figure 10(b), with brown circles representing the original contoured NVBs. Figure 11 provides pseudocode for producing the approximate regions with a combination of transformations. While this region oversimplifies the anatomy surrounding the prostate as a homogenous space created with a combination of a few selected ROIs, analysis of the dose features that are extracted in relation to a patient outcome can provide some insight into the physiological correlations in this region.



(a) (b)

Figure 10: Approximation of neurovascular bundles

(a) Proposed definition of neurovascular regions using a linear combination of contours:

(Prostate + 1.0 cm x,y expansion) – Prostate – (Rectum + 0.2 cm x,y expansion)

Sector boundaries at 30°, 90°, 180°, 270°, and 330°, with 0° pointing anterior to the patient.

(b) Sample of slice showing generated regions. Contoured NVBs (brown) are shown to be encapsulated by the derived sectors.

Helper method: Pull ROIs from database

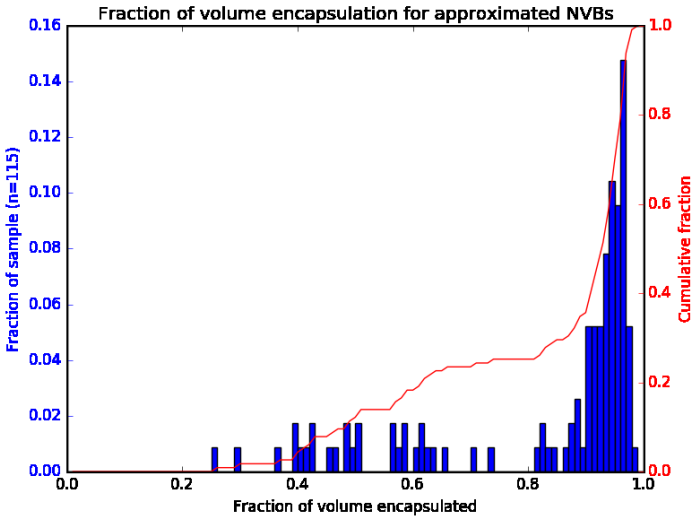
```
function get_masks(patient_rep_id, roi_names):
    '''
    This function queries ROI images for a patient
    Arguments:
        patient_rep_id:  patient identifier
        roi_names:       names of ROIs to query
    Returns:
        masks: dictionary of ROI names to patient images
    '''
    masks = {}
    for roi in roi_names:
        masks[roi] = query_mask_from_database(patient_rep_id, roi)
    return masks
```

Create approximate region using shape transformations

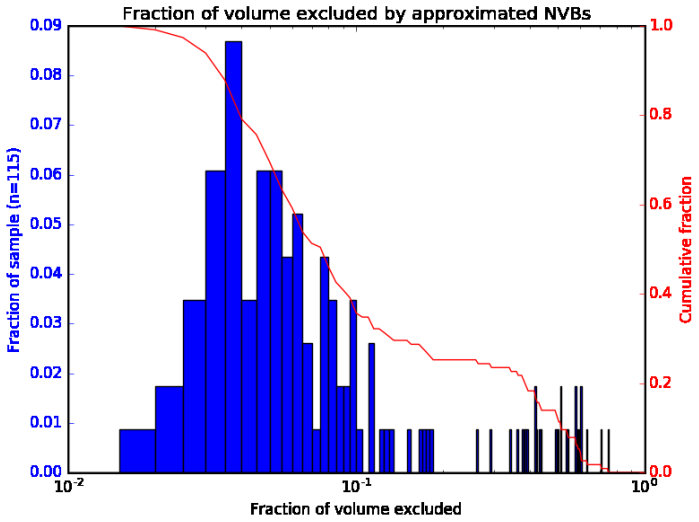
```
import transform

assign patient_rep_id
assign roi_names = ['prostate', 'rectum']
assign prostate_expansion
assign rectum_expansion
assign sector_angles

masks = get_masks(patient_rep_id, roi_names)

# Create the shell around the prostate
step1a = transform.expand(masks['prostate'], prostate_expansion)
step1b = logical_xor(masks['prostate], step1a)

# Remove overlap with rectum
step2a = tf.expand(masks['rectum'], rectum_expansion)
step2b = logical_xor(step_1b, step_2a)

# Create sectors
prostate_sectors = transform.compute_sectors(step2b, sector_angles)
```

Figure 11: Pseudocode for generation of neurovascular regions

Figure 12 validates that the approximate regions accurately captured the NVBs in the contoured patients, showing that 70% of the generated structures encapsulated over 80% of their respective NVB contours.



(a) Differential (blue) and cumulative (red) histograms of fraction of contour volume encapsulated by derived regions.



(b) Differential (blue) and cumulative (red) histograms of fraction of contour volume excluded by derived regions, shown on a logarithmic scale for volume.

Figure 12: Histogram comparing overlap of derived NVB regions and manual contours

4.2 Clinical outcome prediction

The primary goal that drove the development of radio-morphology was clinical outcome prediction: the ability to understand the relationship between radiation dose deposition to the patient anatomy and post-treatment radiation-induced toxicities. The anatomy normalization, shape transformation, and dose feature extraction steps produce a feature set that spatially characterizes the dose to the anatomy of a patient population. The characteristics that are

computed are normalized across the patient set, ensuring that the spatial regions are consistent. Using the dose features alongside patient covariates, such as age, baseline toxicity grade, etc. it is possible to set up a classification problem to predict changes to a patient's QoL following radiotherapy. Figure 13 shows this outcome prediction workflow, using RM feature sets generated at the end of the pipeline in Figure 6.

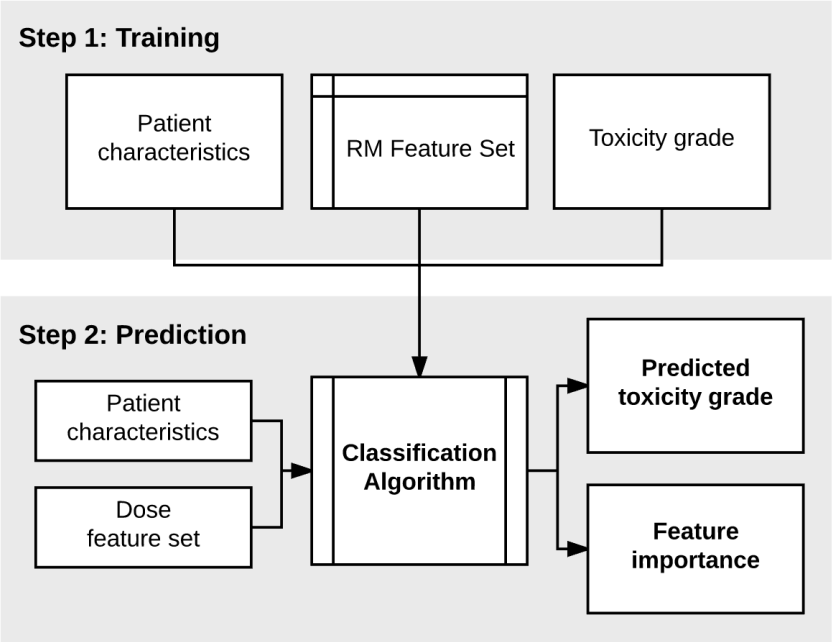


Figure 13: Outcome prediction processing workflow

This application was put into practice to predict the worsening of post-treatment xerostomia in HN cancer patients. Xerostomia refers to the loss of salivary function, which occurs when salivary glands in the HN region are damaged during radiotherapy. It is one of the most widely reported radiation induced toxicities, and it has been shown be linked to overall patient QoL [13].

4.2.1 Voxel-based analysis of parotid and submandibular glands

With a cohort of 427 patients, voxel-based analysis was performed by sampling dose across the parotid and submandibular glands. Each patient's ROIs were sampled and registered to a common image. The set of structures was also normalized by laterality to produce a set of 943 points across the ipsi- and contra-lateral parotid and submandibular glands. Appendix B provides pseudocode for this algorithm.

The 943 dose values sampled from the ROIs and baseline covariates, such as xerostomia grade at the start of treatment and age, were used to train a ridge logistic regression classifier. The ridge classifier was selected due to its ability to identify predictive features and penalize low influence terms to prevent overfitting [14]. The regression method identified regions of high importance across the anatomy; Figure 14 shows the importance of each voxel to the prediction of the high-grade outcome, normalized to a scale of 0-1. The importance patterns show that dose to the superior and medial areas of the contralateral parotid are correlated with the occurrence of high grade xerostomia.

The clinical interpretation of these results is based on the insight of the average dose distribution, shown in Figure 4(a). The average dose distribution shows that the high importance regions that appear on the superior part of the contralateral parotid have a generally lower dose across the patient set. This be interpreted that, when designing a treatment plan, this high importance region is the last area to be irradiated; that is to say, dose to the ipsilateral structures is expected to always be higher, and if this region is receiving a high dose, the overall dose to the patient will similarly be high, resulting in the occurrence of high grade xerostomia. Voxel-based analysis provided an initial understanding of the sub-anatomical differences in an ROI and the effect on the occurrence of negative post-treatment outcomes. The next section

presents a shape-based method for deriving features for the same task and uses knowledge of the sub-morphology of the parotid and glands and the dose statistics in Figure 4 for a similar interpretation.

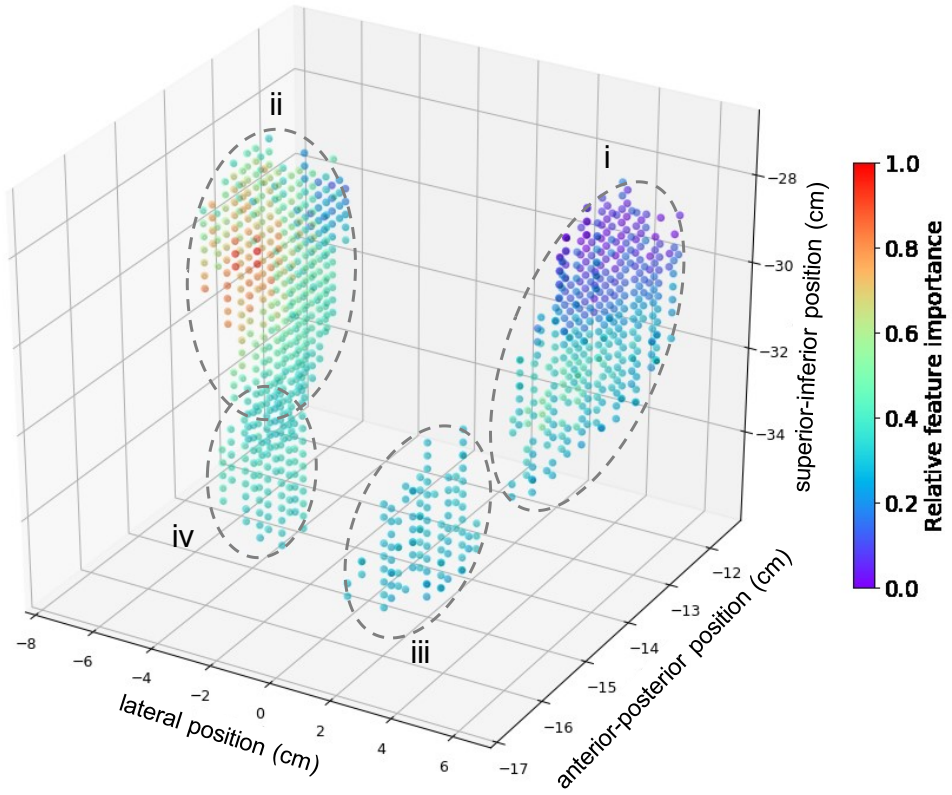


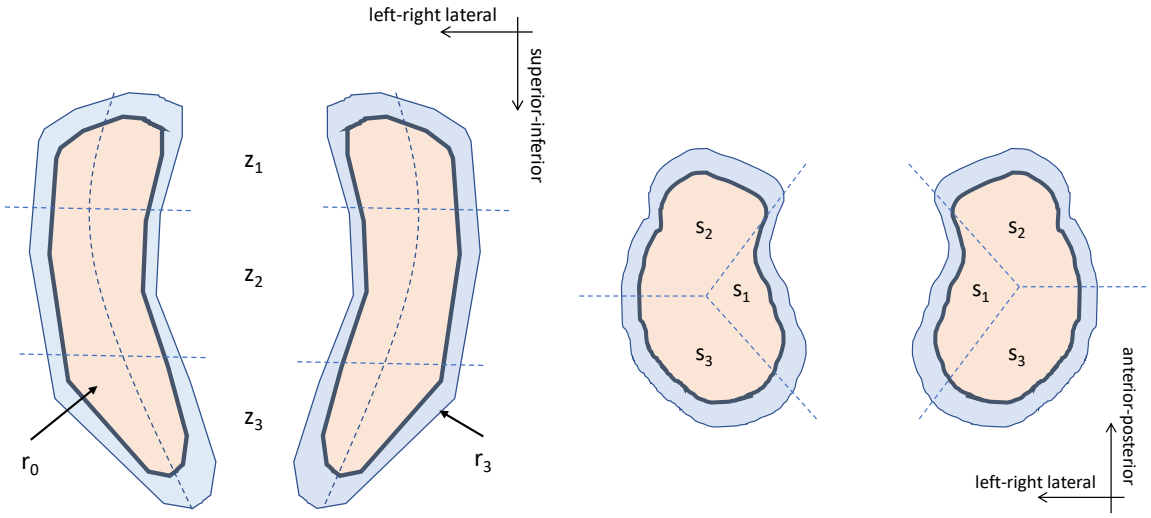
Figure 14: Spatial distribution of feature importance using voxel-based analysis

In this figure, four structures are identified:

i. Ipsilateral parotid gland
ii. Contralateral parotid gland
iii. Ipsilateral submandibular gland
iv. Contralateral submandibular gland

4.2.2 Shape-based DVH features

Following the voxel-based analysis to identify dose patterns that were correlated with the outcome, parametric shape based features were designed to isolate the regions with similar dose patterns, and DVH parameters were calculated for each sub-region. The derived shapes were based on a composition of three transformations: radial expansions to capture a shell surrounding the parotid regions, lateral slices along the superior-inferior axis, and sectors. A shell denoting the space from the surface of the parotid to a 3-mm expansion captures a region of high dose gradient outside the surface of the parotids. The parotid was separated into superior, medial, and inferior slices to isolate differences in influence that were seen in Figure 14. Finally, sectors were defined to be bounded by angles of 45°, 135° and 270° for the left parotid and 90°, 225° and 315° for the right parotid. The combination of the three sets of transformations produces 18 derived structures per parotid gland (36 per patient). The submandibular glands were also considered using the original ROI and the shell of a 3-mm expansion around the ROIs. Based on the laterality of the patient's tumor, the set of structures was reorganized to account for ipsi- and contra-laterality. For each derived region, DVHs were computed and dose-values were sampled in 5% volume increments from 5-95% of the volume of that shape; the resulting dataset consists of 760 dosimetric features per patient. Figure 15 and Table 1 show a cartoon representation of the derived structures with each separation labeled.



(a) Coronal view (b) Axial view

Figure 15: Representation of substructures created from parotid glands

Radial expansions	
r_0	Bounded by surface of parotid
r_3	Shell bounded by 3-mm expansion of parotid and surface of region r_0
Axial slices	
z_1	Superior slice
z_2	Medial slice
z_3	Inferior slice
Sectors	
s_1	Medial sector
s_2	Anterior-lateral sector
s_3	Posterior-lateral sector

Table 1: Regional encoding for substructures in Figure 15

Again, ridge logistic regression was used to learn the DVH features that correlate to the occurrence of high grade xerostomia. This analysis identified that the superior and medial slices, and the anterior-lateral sector of contralateral parotid gland are highly indicative of the development of high grade xerostomia, supporting the findings from the voxel-based analysis. Table 2 lists the ten most important regions and their relative importance when predicting high grade post-treatment xerostomia. Regions are encoded as a combination of the labels from Table 1, "PC" refers to the contralateral parotid, and "dX" refers to the dose delivered to X% of the volume of the structure. Table 3 in Appendix C contains a more extensive version of this table with the first 70 regions in order of relative importance.

Region	Relative importance
PC r_3 s_2 z_2 d80	1.0
PC r_3 s_2 z_2 d85	0.9834919147
PC r_3 s_2 z_2 d75	0.9787498750
PC r_0 s_2 z_1 d95	0.9741924248
PC r_3 s_2 z_2 d70	0.9286902174
PC r_0 s_2 z_1 d90	0.9264683540
PC r_3 s_2 z_2 d65	0.9153294381
PC r_3 s_2 z_2 d55	0.9106901170
PC r_3 s_2 z_1 d95	0.9060323167
PC r_0 s_3 z_1 d75	0.9013485967

Table 2: Relative feature importance in prediction of high grade post-treatment xerostomia.

Considering the sub-morphology of the parotids permits a hypothesis as to why the selected regions are of highly correlated to the occurrence of xerostomia. The salivary ducts in the parotids leave the structure from the medial-to-superior portion of the anterior face, which are the same regions identified by the RM feature analysis [15]. Dissection of the parotid glands have also identified regions of stem cells along the duct, which can contribute to recovery of parotid gland function post radiotherapy treatment [16]. It is then hypothesized that the ductal

region of the parotids acts as a serial component, and damage to this region prevents the flow of saliva out of the parotids, resulting in high grade xerostomia.

The value of the RM pipeline is shown in this set of applications, where it was possible to discern a spatial relationship between radiation dose and clinical outcomes, and identify functionally important regions of the anatomy. Since derived shapes and features can be tuned by modifying the parameters used to create them, predictive models can be optimized by changing the shapes to maximize predictive power and thus, discover regions of the anatomy that are linked to function.

5 Conclusions

Radio-morphology was presented as a method to characterize radiation dose to a patient in an anatomically and physiologically motivated manner. Through the steps of anatomy normalization, image transformation, and dose feature extraction, it is possible to better describe a treatment without masking spatially encoded functional relationships. The feature generation pipeline described each of the major steps and provided examples of implementations at each phase, such as the use of CPD as a non-rigid registration algorithm for point sets, rather than image based registration. The parametric definition of shapes with scaling and partitioning transformations affords a function that is valuable to machine learning applications where features can be modified and tuned simply by adjusting the parameters. This parameterization produces a quantitative representation of human anatomy, allowing for consistent and mathematically motivated exploration of the physiological relationships between organs and their response to radiation.

Several applications of the RM framework were explored to support the merit of this technique. The anatomy normalization step allows for population wide dose statistics, such as the average spatial distribution of radiation across a set of anatomy during therapy. Applying the methodology to analyze spatial dose relationships in HN cancer treatment isolated dose to regions of the contralateral parotid gland as indicative of the occurrence of high grade post-treatment xerostomia. The regions identified through voxel-based and shape-based analysis support the hypotheses that: 1) damage to the lateral section of the contralateral parotid only occurs after high overall dose to all salivary organs, and 2) damage to the parotid ducts found along the anterior face in the medial-to-superior regions prevents salivary function from the rest of the structure.

The growth of big data in medicine has made a large amount of information available; in radiotherapy, this increased scale of data has made it possible to uncover new knowledge about the human body and methods to design better treatments. The RM feature generation pipeline is an effective method for characterizing dose and reveals applications and insight into clinical data that were previously time-consuming or even impossible.

Appendix A `oncotools` Python Library

This appendix contains selected excerpts of code from the `oncotools` Python library described in Section 3.2. Full source code and documentation for the library can be found at:

`https://svn.radonc.jhmi.edu/svn/Oncospace/trunk/AnalyticsFramework`

The `Feature` abstract class

`oncotools/features/feature.py`

```
'''
Features are consistently identifiable characteristics of images or masks.
'''

from copy import deepcopy

class Feature(object):
    '''
    The Feature class is an abstract representation
    of a region of interest's features.

    Each subclass can take different arguments and perform different processing steps.

    Positional arguments:
        :patID: patient identifier
        :roi:   name of the region of interest
    Properties:
        :patientID:         patient identifier
        :roi_name:          name of the region of interest
        :feature_type:      identifier for the type of feature calculation
        :loaded:            flag for whether or not necessary values have been loaded
    Loaded properties:
        :feature_mask:      binary mask that is loaded
        :feature_dose:      dose grid that is loaded
    Calculated properties:
        :feature_dosemaps:  list of dose_masks making up the feature
        :values:            the values associated with the feature
    '''
    def __init__(self, patID, roi):
        self.patientID = patID
        self.roi_name = roi
        self.feature_type = None
        self.feature_dosemaps = None
        self.output = None
        self.loaded = False
        self.feature_dose = None
        self.feature_mask = None
```

```
@property
def values(self):
    '''
    Get the values.

    Each type of feature class has a different type of output.
    See each class's documentation for further details.
    '''
    if self.output is None:
        self.process()
    return self.output

@property
def dosemaps(self):
    '''
    Get the dose maps.
    '''
    if self.feature_dosemaps is None:
        self.process()
    return self.feature_dosemaps

def load(self, feature_mask, feature_dose):
    '''
    Load the feature with a mask and a dose grid.

    Positional arguments:
        :feature_mask:   mask to be processed
        :feature_dose:   dose grid
    '''
    self.feature_mask = deepcopy(feature_mask)
    self.feature_dose = deepcopy(feature_dose)
    self.loaded = True

def process(self):
    '''
    Process the data to calculate the values.

    The process() method is different for each type of feature.
    '''
    raise NotImplementedError("Feature processing not implemented")
```

The `VolumetricFeature` class

`oncotools/features/volumetric_feature.py`

```
'''
The VolumetricFeature class is an extension of the Feature abstract class.

This class uses a list of expansions and contractions
to create a set of concentric shells.
Values are calculated as the average dose between in each shell
and specified DVH values.
'''

from oncotools.utils import transform as tf
from oncotools.utils.dose_map import DoseMask
from oncotools.features.feature import Feature

class VolumetricFeature(Feature):
    '''
    The VolumetricFeature class is an extension of the Feature abstract class.

    This class uses a list of expansions and contractions
    to create a set of concentric shells.
    Values are calculated as the average dose between in each shell
    and specified DVH values.

    Keyword arguments:
        :contract:  list of contractions to be performed on the mask
        :expand:    list of expansions to be performed on the mask
        :dvh:       list of dvh volumes to look up

    Note: Contraction and Expansion values may be given as:
        - A single number:       uniform expansion in all dimensions
        - A list of 1 number:    uniform expansion in all dimensions
        - A list of 3 numbers:   expansion in x, y, and z, respectively
        - A list of 6 numbers:   expansion in -x, +x, -y, +y, -z, and +z, respectively

    Values:
        The values field of a VolumetricFeature is a tuple of lists where...
            - the first element is a list of expansion and contraction factors,
            ordered to define the outer bounds of each shell
            - the second element is a list of average dose in each shell

        Note:
            Given `n` total expansions and contractions, there will be `n+1` dose values.
    '''

    def __init__(self, patID, roi, contract=[], expand=[], dvh=[]):
        if len(contract) == 0 and len(expand) == 0:
            raise ValueError(
                "VolumetricFeature must have either contraction or expansion parameters")
        super(VolumetricFeature, self).__init__(patID, roi)
        self.feature_type = 'VolumetricFeature'
        self.contractions = contract
        self.expansions = expand
        self.dvh = dvh
```

```
def process(self):
    '''
    Process the feature and calculate the average dose in each shell.

    Returns: Dictionary with the following keys:
        :bounds:    List of expansion and contraction factors,
                        ordered to define the bounds of each shell
        :avg:       list of bounds, list of average dose per shell),
        :dvh:       list of volumes, list of dose to volume
    Raises:
        :ValueError:    if data is not loaded
    '''
    if not self.loaded:
        raise ValueError("Load data before processing")
    # Create shells using given contractions and expansions
    bounds, shls = tf.shells(
        self.feature_mask, contractions=self.contractions, expansions=self.expansions)
    # Calculate the dose masks for each shell
    self.feature_dosemaps = [DoseMask(m, self.feature_dose) for m in shls]
    '''
    Output is:
    {
        "avg" : (list of bounds, list of average dose per shell),
        "dvh" : (list of volumes, list of dose to volume)
    }
    '''
    self.output = {
        "bounds": bounds,
        "avg": [dm.average_dose() for dm in self.feature_dosemaps],
        "min": [dm.min_dose() for dm in self.feature_dosemaps],
        "max": [dm.max_dose() for dm in self.feature_dosemaps]
    }

    # If there are specified dvh paramaters, look thmm up
    if len(self.dvh) > 0:
        self.output["dvh"] = (self.dvh, [[dm.get_dose_to_volume(v) for v in self.dvh]
                                         for dm in self.feature_dosemaps])

    return self.output
```

Appendix B Pseudocode for voxel-based analysis

Function: Pull ROIs from database

```
function get_masks(patient_rep_id, roi_names):
    '''
    This function queries ROI images for a patient
    Arguments:
        patient_rep_id:  patient identifier
        roi_names:       names of ROIs to query
    Returns:
        masks: map of ROI names to patient images
    '''
    masks = {}
    for roi in roi_names:
        masks[roi] = query_mask_from_database(patient_rep_id, roi)
    return masks
```

Function: Find point correspondences between two sides

```
function find_corr(l_img, r_img):
    '''
    This function uses CPD to establish point correspondences
    Arguments:
        l_img:  Left image
        r_img:  Right image
    Returns:
        corr:   Compute pairs of points correspondences between left and
                right lateral structures
    '''
    # Mirror one image across x axis
    mirror_img = r_img*[-1, 1, 1]
    # Use CPD to register the two point clouds
    reg_img = register(mirror_img, l_img)
    # Find index of closest point
    corr = [[index_of_closest_point(l_img, pt), i]
             for i, pt in enumerate(reg_img)]
    return corr
```

Function: Reorder sets of points based on point correspondences

```
function reindex_points(l_img, r_img, corr, lat):
    '''
    This function uses CPD to establish point correspondences
    Arguments:
        l_img:  Left image
        r_img:  Right image
        corr:   List of point correspondences
        lat:    Laterality (left vs right)
    Returns:
        ipsi:   Reorganized points on ipsilateral structure
        ctra:   Reorganized points on contralateral structure

    '''
    left_reorg = [l_img[pair[0]] for pair in corr]
    right_reorg = [r_img[pair[1]] for pair in corr]
    if lat == 'left':
        ipsi = left_reorg; ctra = right_reorg;
    else if lat == 'right':
        ipsi = right_reorg; ctra = left_reorg;
    return ipsi, ctra
```

Sample dose to voxels across parotid and submandibular glands

```
import transform
import DoseMask

given atlases = {
    'l_parotid': left_parotid_atlas,
    'r_parotid': right_parotid_atlas
    'l_submandibular': left_submandibular_atlas,
    'r_submandibular': right_submandibular_atlas
}

assign patient_rep_id
assign roi_names = ['l_parotid', 'r_parotid',
                    'l_submandibular', 'r_submandibular']

given laterality

# Query masks and dose from database
masks = get_masks(patient_rep_id, roi_names)
dose = query_dose_from_database(patient_rep_id)

# For each ROI, sample dose points
features = {}
for roi_name in atlases:
    cpd_img = register(atlases[roi_name], masks[roi_name])
    dm = DoseMask(cpd_img, dose)
    features[roi_name] = dm.data

# Reorder features based on point correspondences
corr_parotid = find_corr(features['l_parotid'], features['r_parotid'])
ipsi_par, ctra_par = reindex_points(features['l_parotid'], features['r_parotid']
                                    corr_parotid, laterality)

corr_submand = find_corr(features['l_submandibular'], features['r_submandibular'])
ipsi_sub, ctra_sub = reindex_points(features['l_submandibular'],
                                    features['r_submandibular']
                                    corr_submand, laterality)

# Organize the dose features as one list
feature_list = [ipsi_par, ctra_par, ipsi_sub, ctra_sub]
return feature_list
```

Appendix C Data Tables

Region	Relative importance	Region	Relative importance
PC r3 s2 z2 d80	1	PC r3 s2 z2 d35	0.830390361
PC r3 s2 z2 d85	0.983491915	PC r0 s3 z2 d90	0.830220935
PC r3 s2 z2 d75	0.978749875	PC r0 s2 z1 d75	0.822770383
PC r0 s2 z1 d95	0.974192425	PC r0 s2 z2 d40	0.816006363
PC r3 s2 z2 d70	0.928690217	PC r0 s3 z1 d54	0.812510167
PC r0 s2 z1 d90	0.926468354	PC r0 s2 z1 d80	0.809841264
PC r3 s2 z2 d65	0.915329438	PC r0 s2 z2 d35	0.805722827
PC r3 s2 z2 d54	0.910690117	PC r3 s2 z1 d90	0.805086871
PC r3 s2 z1 d95	0.906032317	PC r3 s3 z1 d85	0.801532183
PC r0 s3 z1 d75	0.901348597	PC r3 s2 z2 d95	0.800268194
PC r3 s2 z2 d49	0.899073082	PC r0 s3 z1 d49	0.795761560
PC r3 s2 z2 d60	0.888584585	PC r3 s2 z2 d30	0.794925779
PC r0 s3 z1 d80	0.888366943	PC r3 s3 z1 d90	0.785938418
PC r0 s2 z2 d85	0.887692036	PC r0 s2 z1 d70	0.780655701
PC r0 s3 z1 d70	0.882679576	PC r0 s3 z2 d85	0.778791433
PC r0 s3 z2 d95	0.882583817	PC r3 s3 z1 d80	0.772052690
PC r0 s3 z1 d85	0.881098447	PC r3 s3 z1 d95	0.767640859
PC r3 s2 z2 d90	0.876432178	PC r0 s2 z2 d30	0.766293920
PC r0 s2 z2 d90	0.875152987	PC r0 s3 z1 d44	0.759773713
PC r0 s2 z1 d85	0.873516176	PC r3 s3 z1 d35	0.758370786
PC r3 s2 z2 d44	0.870748783	PC r3 s3 z1 d40	0.754800822
PC r0 s2 z2 d80	0.864522055	PC r3 s3 z1 d30	0.753868623
PC r3 s2 z2 d40	0.864295610	PC r3 s3 z1 d75	0.747250904
PC r0 s2 z2 d75	0.864156877	PC r3 s3 z1 d44	0.745031957
PC r0 s3 z1 d65	0.862682642	PC r0 s3 z2 d80	0.744792355
PC r0 s2 z2 d60	0.862308302	PC r0 s3 z1 d40	0.744757477
PC r0 s2 z2 d70	0.861594646	PC r0 s3 z1 d35	0.743609599
PC r0 s2 z2 d54	0.858358966	PC r3 s2 z2 d25	0.739874772
PC r0 s3 z1 d90	0.855841448	PC r3 s3 z1 d70	0.737661939
PC r0 s2 z2 d65	0.854652860	PC r0 s2 z2 d25	0.731790041
PC r0 s2 z2 d95	0.851876919	PC r3 s3 z1 d65	0.728062351
PC r0 s3 z1 d95	0.847720692	PC r0 s2 z1 d65	0.724069718
PC r0 s3 z1 d60	0.845766347	PC r3 s2 z1 d85	0.721657385
PC r0 s2 z2 d44	0.837189430	PC r3 s3 z1 d54	0.721625439
PC r0 s2 z2 d49	0.834683694	PC r3 s3 z1 d60	0.717217070

Table 3: Top 70 shape-based features for prediction of high grade post-treatment xerostomia

References

[1] R. Bellazzi, "Big Data and Biomedical Informatics: A Challenging Opportunity," *Yearbook of Medical Informatics,* pp. 8-13, 9 2014.

[2] T. R. McNutt, K. L. Moore and H. Quon, "Needs and Challenges for Big Data in Radiation Oncology," *International Journal of Radiation Oncology,* vol. 94, no. 3, pp. 905-915, 1 July 2016.

[3] K. K. Brock, S. Mutic, T. R. McNutt, H. Li and M. L. Kessler, "Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132," *Medical Physics,* vol. 44, no. 7, pp. 43-76, 23 May 2017.

[4] S. Monti, G. Palma, V. D'Avino, M. Gerardi, G. Marvaso, D. Ciardo, R. Pacelli, B. A. Jereczek-Fossa, D. Alterio and L. Cella, "Voxel-based analysis inveils regional dose differences associated with radiation-induced morbidity in head and neck cancer patients," *Scientific Reports,* 3 August 2017.

[5] A. Myronenko and X. Song, "Point Set Registration: Coherent Point Drift," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 12, pp. 2262-2275, December 2010.

[6] T. McNutt, K. Evans, J. Moore, W. Yang, J. Herman, H. Quon, A. Sharabi, J. Wong and T. DeWeese, "Oncospace: A Database Designed for Personalized Medicine in Radiation Oncology," in *Medical Physics*, Indianapolis, 2013.

[7] The Johns Hopkins University, Department of Radiation Oncology, "Oncospace - Schema Overview," 2007. [Online]. Available: https://oncospace.radonc.jhmi.edu/Schema/Overview.aspx.

[8] Python Software Foundation, "Python," [Online]. Available: https://www.python.org/.

[9] Enthought, Inc., "SciPy," [Online]. Available: https://scipy.org/.

[10] Jupyter, "Project Jupyter," [Online]. Available: http://jupyter.org/.

[11] Microsoft, "Python SQL Driver - pyodbc," Microsoft, 9 August 2017. [Online]. Available: https://docs.microsoft.com/en-us/sql/connect/python/pyodbc/python-sql-driver-pyodbc.

[12] Y. H. Park, C. W. Jeong and S. E. Lee, "A comprehensive review of neuroanatomy of the prostate," *Prostate International,* vol. 1, no. 4, pp. 139-145, 30 December 2013.

[13] A. P. Jellema, B. J. Siotman, P. Doornaert, C. R. Leemans and J. A. Langengjik, "Impact of Radiation-Induced Xerostomia on Quality of Life After Primary Radiotherapy Among Patients With Head and Neck Cancer," *International Journal of Radiation Oncology,* vol. 69, pp. 751-760, 1 November 2007.

[14] S. Le Cessie and J. Van Houwelingen, "Ridge Estimators in Logistic Regression," *Journal of the Royal Statistical Society (Applied Statistics),* vol. 41, no. 1, pp. 191-201, 1992.

[15] A. Osamo, K. Mizobe, Y. Bando and K. Sakiyama, "Anatomy and Histology of Rodent and Human Major Salivary Glands," *Acta Histochemica et Cytochemica,* vol. 45, no. 5, pp. 241-250, 31 October 2012.

[16] P. van Lujik, S. Pringle, J. Deasy, V. Moiseenko, H. Faber, A. Hovan, M. Baanstra, H. van der Laan, R. Kierkels, A. van der Schaaf, M. Witjes, J. Schippers, S. Brandenburg, J. Langendjik, J. Wu and R. Coppes, "Sparing the region of the salivary gland containing stem cells preserves saliva production after radiotherapy for head and neck cancer," *Science Translational Medicine,* vol. 7, no. 305, 16 September 2015.

Biography

Pranav Lakshminarayanan was born in 1994 in Chennai, India, moving to the United States of America in 1999.

Pranav did his undergraduate work at Johns Hopkins University where he majored in Biomedical Engineering with a focus in Biomedical Instrumentation, receiving his Bachelor of Science Degree in 2015. In 2016, he joined the Master of Science and Engineering program for Biomedical Engineering at Johns Hopkins University, researching big data applications in radiation oncology with Todd McNutt, Russell Taylor, and the Oncospace team at the Johns Hopkins Hospital, Department of Radiation Oncology.